

Lower Bounds on Performance of Metric Tree Indexing Schemes for Exact Similarity Search in High Dimensions

Vladimir Pestov

Department of Mathematics and Statistics, University of Ottawa,
585 King Edward Avenue, Ottawa, Ontario K1N 6N5 Canada
e-mail: vpest283@uottawa.ca

The date of receipt and acceptance will be inserted by the editor

Abstract

Within a mathematically rigorous model borrowed from statistical learning theory, we analyse the curse of dimensionality for similarity based information retrieval in the context of popular indexing schemes: metric trees. The datasets X are sampled randomly from a domain Ω , equipped with a distance, ρ , and an underlying probability distribution, μ . While performing an asymptotic analysis, we send the intrinsic dimension d of Ω to infinity, and assume that the size of a dataset, n , grows superpolynomially yet subexponentially in d . Exact similarity search refers to finding the nearest neighbour in the dataset X to a query point $\omega \in \Omega$, where the query points are subject to the same probability distribution μ as datapoints. Let \mathcal{F} denote a class of all 1-Lipschitz functions on Ω that can be used as decision functions in constructing a hierarchical metric tree indexing scheme. Suppose the VC dimension of all sets $\{\omega: f(\omega) \geq a\}$, $a \in \mathbb{R}$ is $d^{O(1)}$. (In view of a result of Goldberg and Jerrum, this is a reasonable complexity assumption.) We deduce superpolynomial in d lower bounds on the expected average case performance of hierarchical metric-tree based indexing schemes for exact similarity search in (Ω, X) .

Introduction

Every similarity query in a dataset with n points can be answered in time $O(n)$ through a simple linear scan of the dataset, and in practice such a scan often outperforms the best known indexing schemes for high-dimensional workloads. This is known as the curse of dimensionality, the cf. e.g. Chapter 9 in [33], as well as [4, 41]. Paradoxically, there is still no mathematical proof that the above phenomenon is in the nature of high-dimensional datasets.

While the concept of intrinsic dimension of data is open to a discussion (see e.g. [30]), even in cases commonly accepted as “high-dimensional” (e.g. uniformly distributed data in the Hamming cube $\{0, 1\}^d$ as $d \rightarrow \infty$), the “curse of dimensionality conjecture” for proximity search remains unproven [15]. Diverse results in this direction [5, 3, 28, 8, 34], are still preliminary.

Here we will verify the conjecture for a particular class of indexing schemes widely used in similarity search and going back to [36]: metric trees. So are called hierarchical partitioning indexing schemes equipped with 1-Lipschitz (non-expanding) decision functions at every node.

We assume that datapoints are drawn from the domain Ω with regard to an underlying probability measure μ independently of each other. The domain is a metric space, that is, the similarity measure, ρ , satisfies the axioms of a metric. The intrinsic dimension of Ω is defined in terms of concentration of measure as in [30]. This concept agrees with the usual notion of dimension in cases such as the Hamming cube $\{0, 1\}^d$ or the Euclidean ball \mathbb{B}^d , and is most relevant. A dataset $X \subseteq \Omega$ with n points is modelled by i.i.d. random variables distributed according to μ . We assume, as in [15], that the number of datapoints n grows superpolynomially in dimension d yet subexponentially in d . Using the notation of asymptotic algorithm analysis, this can be written as $n = d^{\omega(1)}$ and $d = \omega(\log n)$.

It is clear that the computational complexity of decision functions used in constructing a metric tree is a major factor in a scheme performance. We take this into account in the form of a combinatorial restriction on the subclass \mathcal{F} of all functions on Ω that are allowed to be used as decision functions, by requiring a well-known parameter of statistical learning theory, the Vapnik-Chervonenkis dimension of \mathcal{F} [37], to be polynomial in d , that is, $\text{VC-dim}(\mathcal{F}) = d^{O(1)}$.

A very general class of functions satisfying this VC dimension bound is provided by a theorem of Goldberg and Jerrum [12] about function classes parametrized by elements of \mathbb{R}^s whose computation involves arithmetic operations, conditioning on inequalities, and inputs 0 or 1. Apparently, the decision functions of all indexing schemes used in practice so far in Euclidean (and Hamming cube) domains fall into this class.

Under above assumptions, we prove a superpolynomial in d lower bound on the expected average performance of all possible metric trees.

The domains of interest are the Euclidean space \mathbb{R}^d equipped with the gaussian measure γ^d , the cube $[0, 1]^d$ with the uniform measure, the Euclidean sphere \mathbb{S}^n with the Haar (Lebesgue) measure, and the Hamming cube $\{0, 1\}^n$ with the Hamming distance and the counting measure. In order to treat all of these simultaneously, we work in a general setting of metric spaces with measure. This approach, in our view, helps to stress the underlying geometry while ignoring the unnecessary detail.

1 General framework for similarity search

We follow a formalism of [14] as adapted for similarity search [28,31]. A *workload* is a triple $W = (\Omega, X, \mathcal{Q})$, where Ω is the *domain*, whose elements can occur as datapoints and as query points, $X \subseteq \Omega$ is a finite subset (*dataset*, or *instance*), and $\mathcal{Q} \subseteq 2^\Omega$ is a family of *queries*. Answering a query $Q \in \mathcal{Q}$ means listing all datapoints $x \in X \cap Q$.

A (dis)similarity measure on Ω is a function of two arguments $\rho: \Omega \times \Omega \rightarrow \mathbb{R}$, which we assume to be a metric, as in [43]. (Although sometimes one needs to consider more general similarity measures, cf. [11,31].) A *range similarity query centred at* $\omega \in \Omega$ is a ball of radius ε around the query point:

$$Q = \mathcal{B}_\varepsilon(\omega) = \{x \in \Omega: \rho(\omega, x) < \varepsilon\}.$$

Equipped with such balls as queries, the triple $W = (\Omega, \rho, X)$ forms a *range similarity workload*.

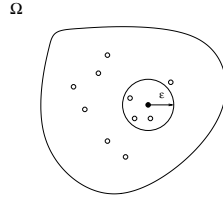


Fig. 1 A range query.

We will assume ρ to be a metric, as in [43],

The *k-nearest neighbours (k-NN) query* centred at $\omega \in \Omega$, where $k \in \mathbb{N}$, is normally being reduced to a range query of a suitable search radius.

A workload is *inner* if $X = \Omega$ and *outer* if $|X| \ll |\Omega|$. There is a substantial difference between the two types of workloads, and most workloads of practical interest are outer workloads, that is, a typical query point will come from outside the dataset, cf. [31].

2 Hierarchical tree index structures

An *access method* is an algorithm that correctly answers every range query. Examples of access methods are given by *indexing schemes*. A hierarchical tree-based indexing scheme is a sequence of refining partitions of the domain labelled with a finite rooted tree. (For simplicity, we will assume all trees to be binary: this is not really restrictive.) Such a scheme takes storage space $O(n)$.

To process a range query $\mathcal{B}_\varepsilon(\omega)$, we traverse the tree recursively to the leaf level. Once a leaf B is reached, its contents (datapoints $x \in X \cap B$) are accessed, and the condition $x \in \mathcal{B}_\varepsilon(\omega)$ verified for each one of them.

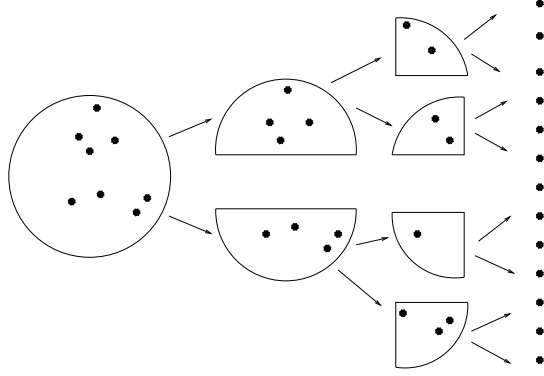


Fig. 2 A refining sequence of partitions of Ω .

Let $C \subseteq \Omega$ be an internal node having child nodes A and B , so that $C = A \cup B$. A branch descending from B can be pruned provided $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, because then datapoints contained in B are of no further interest. This is the case where one can certify that ω is not contained in the ε -neighbourhood of B ,

$$\omega \notin B_\varepsilon = \{x \in \Omega: d(x, B) < \varepsilon\}.$$

(Cf. Fig. 3, l.h.s.) Similarly, if $\omega \notin A_\varepsilon$, then the sub-tree descending from A can be pruned. However, if ω belongs to the intersection of ε -neighbourhoods of A and B , pruning is impossible and the search branches out. (Cf. Fig. 3, r.h.s.)

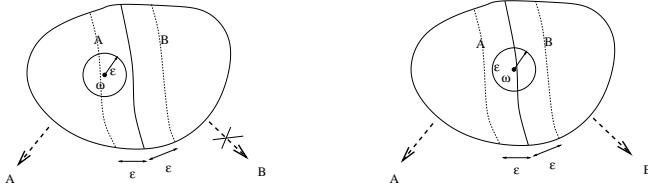


Fig. 3 Pruning is possible (l.h.s.), and impossible (r.h.s.).

In order to certify that $\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset$, one employs *decision functions*. A function $f: \Omega \rightarrow \mathbb{R}$ is a *1-Lipschitz* if

$$\forall x, y \in \Omega, \quad |f(x) - f(y)| \leq d(x, y).$$

Assign to every internal node C a 1-Lipschitz function $f = f_C$ so that $f_C \upharpoonright B \leq 0$ and $f_C \upharpoonright A \geq 0$. It is easily seen that $f_C \upharpoonright B_\varepsilon < \varepsilon$, and so the fact that $\boxed{f_C(\omega) \geq \varepsilon}$ serves as a certificate for $\boxed{\mathcal{B}_\varepsilon(\omega) \cap B = \emptyset}$, assuring that a sub-tree descending from B can be pruned. Similarly, if $f_C(\omega) \leq -\varepsilon$, the sub-tree descending from A can be pruned.

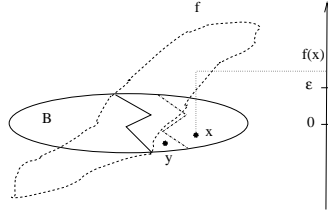


Fig. 4 Graph of a decision function $f = f_C$.

Note that decision functions should have sufficiently low computational complexity in order for the indexing scheme to be efficient.

A hierarchical indexing structure employing 1-Lipschitz decision functions at every node is known as a *metric tree*.

3 Metric trees

Here is a formal definition. A metric tree for a metric similarity workload (Ω, ρ, X) consists of

- a finite binary rooted tree \mathcal{T} ,
- a collection of (possibly partially defined) real-valued 1-Lipschitz functions $f_t: B_t \rightarrow \mathbb{R}$ for every inner node t (decision functions), where $B_t \subseteq \Omega$,
- a collection of *bins* $B_t \subseteq \Omega$ for every leaf node t , containing pointers to elements $X \cap B_t$,

so that

- $B_{\text{root}(\mathcal{T})} = \Omega$,
- for every internal node t and child nodes t_-, t_+ , one has $B_t \subseteq B_{t_-} \cup B_{t_+}$,
- $f_t \upharpoonright B_{t_-} \leq 0$, $f_t \upharpoonright B_{t_+} \geq 0$.

When processing a range query $\mathcal{B}_\epsilon(\omega)$,

- t_- is accessed $\iff f_t(\omega) < \epsilon$, and
- t_+ is accessed $\iff f_t(\omega) > -\epsilon$.

Here is the search algorithm in pseudocode.

Algorithm 1

```

on input  $(\omega, \epsilon)$  do
  set  $A_0 = \{\text{root}(\mathcal{T})\}$ 
  for each  $i = 0, 1, \dots, \text{depth}(\mathcal{T}) - 1$  do
    if  $A_i \neq \emptyset$ 
      then for each  $t \in A_i$  do
        if  $t$  is an internal node
          then do
            if  $f_t(\omega) < \epsilon$ 

```

```

    then  $A_{i+1} \leftarrow A_{i+1} \cup \{t_{-}\}$ 
    if  $f_t(\omega) > -\varepsilon$ 
    then  $A_{i+1} \leftarrow A_{i+1} \cup \{t_{+}\}$ 
  else for each  $x \in B_t$  do
    if  $x \in \mathcal{B}_\varepsilon(\omega)$ 
    then  $A \leftarrow A \cup \{x\}$ 
return  $A$ 
□

```

Under our assumptions on the metric tree, Algorithm 1 correctly answers every range similarity query for the workload (Ω, ρ, X) and thus is an access method.

For more, see [31], while the survey [8] presents a different perspective. Each of the books [32, 33, 43] is an excellent reference to indexing structures in metric spaces.

4 Curse of dimensionality

In practice, a simple linear scan of the dataset, taking time $O(n)$, often outperforms the best known indexing schemes for high-dimensional workloads, though of course there are exceptions, cf. e.g. a relatively efficient scheme developed in [35] for searching large databases of short protein fragments. Nevertheless, in recent years the research emphasis has shifted towards *approximate* similarity search:

- given $\epsilon > 0$ and $\omega \in \Omega$, return a point $x \in X$ that is [with confidence $> 1 - \delta$] at a distance $< (1 + \epsilon)d_{NN}(\omega)$ from ω .

This has led to many impressive achievements, particularly [18, 16], see also the survey [15] and Chapter 7 in [38]. At the same time, research in exact similarity search, especially concerning deterministic algorithms, has slowed down. At a theoretical level, the following unproved conjecture poses a major challenge.

Conjecture 1 (The curse of dimensionality conjecture, cf. [15]) Let $X \subseteq \{0, 1\}^d$ be a dataset with n points, where the Hamming cube $\{0, 1\}^d$ is equipped with the Hamming (ℓ^1) distance:

$$d(x, y) = \#\{i: x_i \neq y_i\}.$$

Suppose $d = n^{o(1)}$, but $d = \omega(\log n)$. (That is, the number of points in X has intermediate growth with regard to the dimension d : it is superpolynomial in d , yet subexponential.) Then any data structure for exact nearest neighbour search in X , with $d^{O(1)}$ query time, must use $n^{\omega(1)}$ space within the *cell probe model* of computation [22].

The best lower bound currently known is $O(d/\log \frac{sd}{n})$, where s is the number of cells used by the data structure [27]. In particular, this implies the earlier bound $\Omega(d/\log n)$ for polynomial space data structures [3], as well as the bound $\Omega(d/\log d)$ for near linear space (namely $n \log^{O(1)} n$). See also [1, 26].

5 Concentration of measure

As in [10], we assume the existence of an unknown probability measure μ on Ω , such that both datapoints X and query points ω are being sampled with regard to μ .

On the one hand, this assumption is open to debate: for instance, in a typical university library most books (75 % or more) are never borrowed a single time, so it is reasonable to assume that the distribution of queries in a large dataset will be skewed equally heavily away from data distribution. On the other hand, there is no obvious alternative way of making an apriori assumption about the query distribution, and in some situations the assumption makes sense indeed, e.g. in the context of a large biological database where a newly-discovered protein fragment has to be matched against every previously known sequence.

The triple (Ω, ρ, μ) is known in a mathematical context as a *metric space with measure*. This concept opens the way to systematically using the *phenomenon of concentration of measure on high-dimensional structures*, also known as the “*Geometric Law of Large Numbers*.” This phenomenon arguably plays at least some role in explaining the nature of the course of dimensionality, and can be informally summarized as follows:

for a typical “high-dimensional” structure Ω , if A is a subset containing at least half of all points, then the measure of the ε -neighbourhood A_ε of A is overwhelmingly close to 1 already for small $\varepsilon > 0$.

Here is a rigorous way for dealing with the phenomenon. Define the *concentration function* α_Ω of a metric space with measure Ω by

$$\alpha_\Omega(\varepsilon) = \begin{cases} \frac{1}{2}, & \text{if } \varepsilon = 0, \\ 1 - \min \{ \mu_\#(A_\varepsilon) : A \subseteq \Omega, \mu_\#(A) \geq \frac{1}{2} \}, & \text{if } \varepsilon > 0. \end{cases}$$

The value of $\alpha_\Omega(\varepsilon)$ gives an upper bound on the measure of the complement to the ε -neighbourhood A_ε of every subset A of measure $\geq 1/2$, cf. Fig. 5.

For example, let $\Omega = \{0, 1\}^d$ be the Hamming cube of dimension d equipped with the normalized Hamming distance

$$d(x, y) = \frac{1}{d} \# \{i : x_i \neq y_i\}$$

and the uniform (normalized counting) measure

$$\mu_\#(A) = \frac{\#A}{2^d}.$$

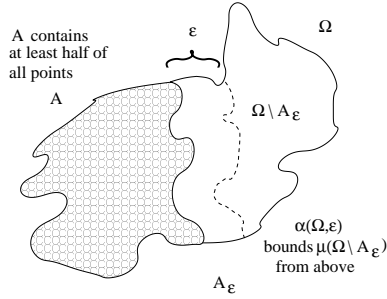


Fig. 5 To the concept of concentration function $\alpha_\Omega(\epsilon)$.

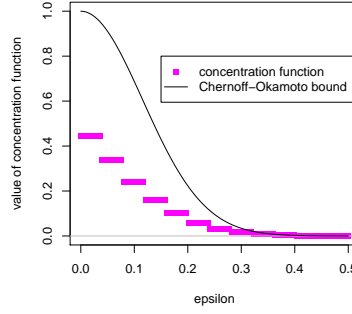


Fig. 6 Concentration function of $\{0, 1\}^{50}$ vs Chernoff–Okamoto bound.

Then the concentration function of Ω satisfies a gaussian upper estimate (Chernoff–Okamoto bound), cf. Fig. 6:

$$\alpha_{\{0,1\}^n}(\epsilon) \leq e^{-3\epsilon^2 d/4}.$$

Similar bounds hold for Euclidean spheres \mathbb{S}^n , cubes \mathbb{I}^n , and many other structures of both continuous and discrete mathematics, equipped with suitably normalized distances and canonical probability measures. The concentration phenomenon can be expressed by saying that for “typical” high-dimensional metric spaces with measure, Ω , the concentration function $\alpha_\Omega(\epsilon)$ drops off sharply as $\dim \Omega \rightarrow \infty$ [21, 19].

6 Workload assumptions

Here are our standing assumptions for the rest of the article. Let (Ω, ρ, μ) be a domain equipped with a metric ρ and a probability measure μ . We assume that the expected distance between two points of Ω is normalized so as to become asymptotically constant:

$$\mathbb{E} \rho(x, y) = \Theta(1). \quad (1)$$

We further assume that Ω has “concentration dimension d ” in the sense that the concentration function α_Ω is gaussian with exponent $\Theta(d)$:

$$\alpha_\Omega(\epsilon) = \exp(-\Theta(\epsilon^2 d)). \quad (2)$$

(This approach to intrinsic dimension is developed in [29, 30].)

A dataset $X \subseteq \Omega$ contains n points, where n and d are related as follows:

$$n = d^{\omega(1)}, \quad (3)$$

$$d = \omega(\log n). \quad (4)$$

In other words, asymptotically n grows faster than any polynomial function Cd^k , $C > 0$, $k \in \mathbb{N}$, but slower than any exponential function e^{cd} , $c > 0$. (An example of such rate of growth is $n = 2^{\sqrt{d}}$.) For the purposes of asymptotic analysis of search algorithms such assumptions are natural [15].

Datapoints are modelled by a sequence of i.i.d. random variables distributed according to the measure μ :

$$X_1, X_2, \dots, X_n \sim \mu.$$

The instances of datapoints will be denoted with corresponding lower case letters x_1, x_2, \dots, x_n .

Finally, the query centres $\omega \in \Omega$ have the same distribution μ :

$$\omega \sim \mu.$$

7 Query radius

As a well-known consequence of concentration, in high-dimensional domains the distance to the nearest neighbour is close to the average distance between two points (cf. e.g. [4] for a particular case). We will give a proof of this result in the most general situation. Denote $\varepsilon_{NN}(\omega)$ the distance from $\omega \in \Omega$ to the nearest point in X . The function ε_{NN} is 1-Lipschitz, and so it concentrates near its median value. From here, one deduces:

Lemma 1 *Under our assumptions on the domain Ω and a random sample X , with confidence approaching 1 one has for all ϵ*

$$\mu \{ \omega : |\varepsilon_{NN}(\omega) - \mathbb{E} \rho(x, y)| > \epsilon \} < \exp(-O(\epsilon^2 d)).$$

□

Remark 1 The result should be understood in the asymptotic sense, as follows. We deal with a family of domains Ω_d , $d \in \mathbb{N}$, and the sampling is performed in each of them in an independent fashion, so that “confidence” refers to the probability that the infinite sample path belonging to the infinite product

$$\Omega_1^{n_1} \times \Omega_2^{n_2} \times \dots \times \Omega_d^{n_d} \times \dots$$

satisfies the desired properties.

The key to the proof is the following technical lemma.

Lemma 2 (Gromov–Milman [13]) *Let (Ω, ρ, μ) denote a metric space with measure and α its concentration function. Then if $A \subseteq \Omega$ is such that $\mu(A) > \alpha(\gamma)$ for some $\gamma > 0$, it implies that $\mu(A_\gamma) > 1/2$. □*

Proof (of Lemma 1) Denote for $\omega \in \Omega$ by ρ_ω the distance function to ω . Being 1-Lipschitz, ρ_ω concentrates near its median value, which we denote by $R(\omega)$. In its turn, the function $R: \Omega \rightarrow \mathbb{R}$ is also 1-Lipschitz, and so concentrates around its median value, R_M . The difference between the mean and the median of every 1-Lipschitz function on Ω converges to zero uniformly as $d \rightarrow \infty$, and for this reason we can substitute R_M for $\mathbb{E}_{\mu \otimes \mu}(\rho)$ in the rest of the proof: indeed, $E_M \rightarrow \mathbb{E}_{\mu \otimes \mu}(\rho) = 1$ in the limit $d \rightarrow \infty$.

Denote ε_M the median value of the nearest neighbour radius ε_{NN} . Suppose $\liminf_{d \rightarrow \infty} \varepsilon_M < 1$. By proceeding to a subsequence of domains, we can select a $\gamma > 0$ so that $\varepsilon_M \leq R_M - \gamma$. If d is sufficiently large, the probability that the function $R(\omega)$ deviates from R_M by more than $\gamma/2$ is exponentially small. Since $n = |X|$ only grows subexponentially in d , with confidence $1 - \exp(-O(\epsilon^2 d))$ one has for every $x \in X$:

$$R(x) - \varepsilon_M \geq \frac{\gamma}{2}.$$

According to Lemma 2,

$$\mu(B_{\varepsilon_M}(x)) \leq \alpha(\gamma/2) = \exp(-O(\epsilon^2 d)),$$

and so

$$\mu(X_{\varepsilon_M}) \leq n \exp(-O(\epsilon^2 d)) = \exp(-O(\epsilon^2 d)),$$

a contradiction, since by the definition of ε_M one must have

$$\mu(X_{\varepsilon_M}) \geq \frac{1}{2}.$$

We conclude: $\liminf_{d \rightarrow \infty} \varepsilon_M \geq 1$.

To establish the converse inequality, suffices to notice that a ball of radius $R(\omega)$ has measure $\geq 1/2$, and so if $\liminf_{d \rightarrow \infty} \varepsilon_M$ is strictly greater than 1, there will be, with high confidence, points $x \in X$ for which $\varepsilon_M > R(x)$ either, and due to concentration, $\mu(X_{\varepsilon_M})$ will be close to one, meaning the complement to the $R(x)$ -ball is very small, in contradiction to the choice of $R(x)$ as the median. \square

8 A “naive” average $O(n)$ lower bound

As a first approximation to our analysis, we present a (flawed) heuristic argument, allowing *linear* in n asymptotic lower bounds on the search performance of a metric tree.

First of all, what happens at an internal node C when a metric tree is being traversed? Let α_C denote the concentration function of C equipped with the metric induced from Ω and a probability measure μ_C which is the normalized restriction of the measure μ from Ω :

$$\text{for } A \subseteq C, \quad \mu_C(A) = \frac{\mu(A)}{\mu(C)}.$$

Suppose for the moment that our tree is perfectly balanced, in the sense that $\mu_C(A) = \mu_C(B) = \frac{1}{2}$. Then the size of the ε -neighbourhood of A is at least $1 - \alpha_C(\varepsilon)$, and the same is true of the ε -neighbourhood of B . One concludes: for all query points $\omega \in C$ except a set of measure $\leq 2\alpha_C(\varepsilon)$, the search algorithm 1 branches out at the node C . (Cf. Fig. 7.)

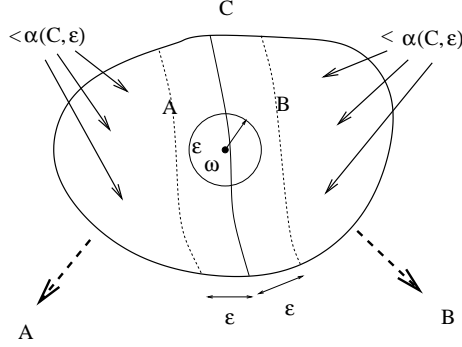


Fig. 7 Search algorithm branches out for most query points ω at a node C if the value $\alpha_C(\varepsilon)$ is small.

Lemma 3 *Let C be a subset of a metric space with measure (Ω, ρ, μ) . Denote α_C the concentration function of C with regard to the induced metric $\rho \upharpoonright C$ and the induced probability measure $\mu/\mu(C)$. Then for all $\varepsilon > 0$*

$$\alpha_C(\varepsilon) \leq \frac{\alpha_\Omega(\varepsilon/2)}{\mu(C)}.$$

Proof Let $\varepsilon > 0$ be any, and let $\delta < \alpha_C(\varepsilon)$. Then there are subsets $D, E \subseteq C$ at a distance $\geq \varepsilon$ from each other, satisfying $\mu(D) \geq \mu(C)/2$ and $\mu(E) \geq \delta\mu(C)$, in particular the measure of either set is at least $\delta\mu(C)$. Since the $\varepsilon/2$ -neighbourhoods of D and E in Ω cannot meet by the triangle inequality, the complement, F , to at least one of them, taken in Ω , has the property $\mu(F) \geq 1/2$, while $\mu(F_{\varepsilon/2}) \leq 1 - \delta\mu(C)$, because $F_{\varepsilon/2}$ does not meet one of the two original sets, D or E . We conclude: $\alpha_\Omega(\varepsilon/2) \geq \delta\mu(C)$, and taking suprema over all $\delta < \alpha_C(\varepsilon)$,

$$\alpha_\Omega(\varepsilon/2) \geq \alpha_C(\varepsilon)\mu(C),$$

that is, $\alpha_C(\varepsilon) \leq \alpha_\Omega(\varepsilon/2)/\mu(C)$, as required. \square

Since the size of the indexing scheme is $O(n)$, a typical size of a set C will be on the order $\Omega(n^{-1})$, while $\alpha_\Omega(\varepsilon)$ will go to zero as $o(n^{-1})$.

Let a workload (Ω, ρ, X) be indexed with a balanced metric tree of depth $O(\log n)$, having $O(n)$ bins of roughly equal size in the sense of the probability measure μ underlying the datapoint distribution.

For at least half of all query points, the distance ε_{NN} to the nearest neighbour in X is at least as large as ε_M , the median NN distance. Let ω be such a query centre. For every element C of level t partition of Ω , one has, using Lemmas 3 and 1 and the assumption in Eq. (2),

$$\alpha_C(\varepsilon_M) \leq \frac{\alpha_\Omega(\varepsilon_M/2)}{\mu(C)^{-1}} = \Theta(2^t)e^{-\Theta(1)\varepsilon_M^2 d} = e^{-\Theta(d)},$$

where the constants *do not depend* on a particular internal node C . An argument in Section 7 implies that branching *at every internal node* occurs for all ω except a set of measure

$$\leq \#(\text{nodes}) \times 2 \sup_C \alpha_C(\varepsilon) = O(n^2)e^{-\Theta(d)} = o(1),$$

because $d = \omega(\log n)$ and so $e^{\Theta(d)}$ is superpolynomial in n . Thus, the expected average performance of an indexing scheme as above is linear in n .

There are two problems with this argument. Firstly, it has been observed and verified experimentally that unbalanced metric tree indexing schemes are more efficient than the balanced ones [7, 24]. Secondly and more importantly, in our argument we have replaced the value of the *empirical measure*,

$$\mu_n(C) = \frac{|C|}{n},$$

with the value of the underlying measure $\mu(C)$, implicitly assuming that the two are close to each other:

$$\mu_n(C) \approx \mu(C).$$

But the scheme is being chosen *after* seeing an instance X , and it is reasonable to assume that the choice of indexing partitions will take advantage of large random clusters always present in uniformly distributed data. (Fig. 8 illustrates this point in dimension $d = 2$.) Thus, some elements of indexing partitions, while having large measure μ , may contain few datapoints, and vice versa.

In order to be able to estimate the empirical measure in terms of the underlying distribution, one needs to invoke an approach of statistical learning.

9 Vapnik–Chervonenkis theory

Let \mathcal{A} be a family of subsets of a set Ω (a *concept class*). One says that a subset $B \subseteq \Omega$ is *shattered* by \mathcal{A} (cf. Fig. 9) if for each $C \subseteq B$ there is $A \in \mathcal{A}$ such that

$$A \cap B = C.$$

The *Vapnik–Chervonenkis dimension* $\text{VC-dim}(\mathcal{A})$ of a class \mathcal{A} is the largest cardinality of a set $B \subseteq \Omega$ shattered by \mathcal{A} .

Estimating the VC dimension is often non-trivial, and here are some examples.

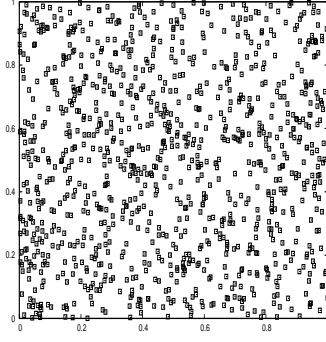


Fig. 8 1000 points randomly and uniformly distributed in the square $[0, 1]^2$.

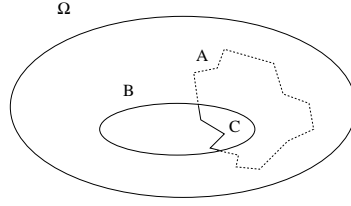


Fig. 9 A set B is shattered by the class \mathcal{A} .

1. The VC dimension of the class of all Euclidean balls in \mathbb{R}^d is $d + 1$.
2. The class of all parallelepipeds in \mathbb{R}^d has VC dimension $2d + 2$.
3. The VC dimension of the class of all ℓ^1 -balls in the Hamming cube $\{0, 1\}^d$ is bounded from above by $d + \lfloor \log_2 d \rfloor$.

(As every ball is determined by its centre and radius, the total number of pairwise different balls in $\{0, 1\}^d$ is $d2^d$. Now one uses an obvious observation: the VC dimension of a finite concept class \mathcal{A} is bounded above by $\log_2 |\mathcal{A}|$.)

Here is a deeper and very general observation.

Theorem 2 (Goldberg and Jerrum [12]) *Consider the parametrized class*

$$\mathcal{F} = \{x \mapsto f(\theta, x) : \theta \in \mathbb{R}^s\}$$

for some $\{0, 1\}$ -valued function f . Suppose that, for each input $x \in \mathbb{R}^n$, there is an algorithm that computes $f(\theta, x)$, and this computation takes no more than t operations of the following types:

- the arithmetic operations $+$, $-$, \times and $/$ on real numbers,
- jumps conditioned on $>$, \geq , $<$, \leq , $=$, and \neq comparisons of real numbers, and
- output 0 or 1.

Then $VC\text{-dim}(\mathcal{F}) \leq 4s(t + 2)$. \square

Now, a typical result of statistical learning theory, where we skip over the measurability assumptions (see [2, 37, 39, 20] for details and more).

Theorem 3 *Let $\mathcal{A} \subseteq 2^\Omega$ be a concept class of finite VC dimension, d . Then for all $\epsilon, \delta > 0$ and every probability measure μ on Ω , if n datapoints in X are drawn randomly and independently according to μ , then with confidence $1 - \delta$*

$$\forall A \in \mathcal{A}, \quad \left| \mu(A) - \frac{|X \cap A|}{n} \right| < \epsilon,$$

provided n is large enough:

$$n \geq \frac{128}{\epsilon^2} \left(d \log \left(\frac{2e^2}{\epsilon} \log \frac{2e}{\epsilon} \right) + \log \frac{8}{\delta} \right).$$

Let \mathcal{F} be a class of (possibly partially defined) real-valued functions on Ω . Define \mathcal{F}_\geq as the family of all sets of the form

$$\{\omega \in \text{dom } f : f(\omega) \geq a\}, \quad a \in \mathbb{R}.$$

The value of $\text{VC-dim}(\mathcal{F}_\geq)$ is bounded above by the Pollard dimension (pseudodimension) of \mathcal{F} (cf. [39], 4.1.2), but is in general smaller.

Example 1 (Pivots) If \mathcal{F} is the class of all distance functions to points of \mathbb{R}^d , then $\text{VC-dim}(\mathcal{F}_\geq) = d + 1$. (The family \mathcal{F}_\geq consists of complements to open balls, and the VC dimension is invariant under proceeding to the complements.) For the Hamming cube, one has $\text{VC-dim}(\mathcal{F}_\geq) \leq d + \lceil \log_2 d \rceil$, but we do not know an exact bound.

Example 2 (vp-tree) The *vp-tree* [42] uses decision functions of the form

$$f_t(\omega) = (1/2)(\rho(x_{t+}, \omega) - \rho(x_{t-}, \omega)),$$

where t_\pm are two children of t and x_{t_\pm} are the *vantage points* for the node t .

If $\Omega = \mathbb{R}^d$, then \mathcal{F}_\geq consists of all half-spaces, and the VC dimension of this family is well known to equal $d + 1$ (in consequence of the classical Radon theorem in convex geometry, just like the proof for Euclidean spheres).

Example 3 (M-tree) The *M-tree* [9] employs decision functions

$$f_t(\omega) = \rho(x_t, \omega) - \sup_{\tau \in B_t} \rho(x_t, \tau),$$

where B_t is a block corresponding to the node t , x_t is a datapoint chosen for each node t , and suprema on the r.h.s. are precomputed and stored.

Here the VC dimension estimates are as in Example 1.

10 Rigorous lower bounds

In this Section we prove the following theorem under general assumptions of Section 6.

Theorem 4 *Let the domain Ω equipped with a metric ρ and probability measure μ have concentration dimension $\Theta(d)$ (cf. Eq. (2)) and expected distance between two points $\mathbb{E}d(x, y) = 1$. Let \mathcal{F} be a class of all 1-Lipschitz functions on the domain Ω that can be used as decision functions for metric tree indexing schemes of a given type. Suppose $VC\text{-dim}(\mathcal{F}_{\geq})$ is polynomial in d . Let X be an i.i.d. random sample of Ω according to μ , having n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$. Then, with confidence asymptotically approaching 1, an optimal metric tree indexing scheme for the similarity workload (Ω, ρ, X) has expected average performance $d^{\omega(1)}$. In other words, the average search time for an exact nearest neighbour is superpolynomial in dimension d .*

The following is an immediate consequence of Lemma 4.2 in [28].

Lemma 4 (“Bin Access Lemma”) *Let $\varepsilon > 0$ and $m \geq 4$ be such that $\alpha_{\Omega}(\varepsilon) \leq m^{-1}$, and let γ be a collection of subsets $A \subseteq \Omega$ of measure $\mu(A) \leq m^{-1}$ each, satisfying $\mu(\cup \gamma) \geq 1/2$. Then the 2ε -neighbourhood of every point $\omega \in \Omega$, apart from a set of measure at most $\frac{1}{2}m^{-\frac{1}{2}}$, meets at least $\frac{1}{2}m^{\frac{1}{2}}$ elements of γ .*

Here is the next step in the proof.

Lemma 5 *Denote \mathcal{B} the class of all subsets $B \subseteq \Omega$ appearing as bins of metric trees of depth $\leq h$ built using certification functions from a class \mathcal{F} satisfying $VC\text{-dim}(\mathcal{F}_{\geq}) \leq p$. Then*

$$VC\text{-dim}(\mathcal{B}) \leq 4hp \log(2hp).$$

Proof Every $B \in \mathcal{B}$ is an intersection of a family of $\leq 2h$ sets from \mathcal{F}_{\geq} or their complements. Now one uses Th. 4.5 in [39]: if \mathcal{A} is a concept class of VC dimension $\leq p$, then the VC dimension of the class of all sets obtained as intersections of $\leq h$ sets from \mathcal{F} is bounded by $2hp \log(hp)$. \square

Let us prove Theorem 4. Without loss in generality, suppose that for any fixed value $0 < c < 1$ such as e.g. $c = 1/4$, for all points ω except in a set of measure $\leq c$ the depth of the search tree is polynomial in d , uniformly in ω , for otherwise there is nothing to prove.

Using Eq. (1) and Lemma 1, pick any $\varepsilon' > 0$ such that, for sufficiently high values of d , for most points ω the value of $\varepsilon_{NN}(\omega)$ exceeds ε' . Let $0 < \beta < 1/2$. Again without losing generality, we can assume that the measure of the set of query centres ω whose ε' -neighbourhood meets at least one bin with $\geq n^{1/2-\beta}$ points is $\leq 1/4$.

Combining the two assumptions together, we deduce that for at least half of all query centres ω the ε' -ball around ω only meets bins with fewer than $n^{1/2-\beta}$ points. By Theorem 3 and Lemma 5, the value of measure μ for each of these bins is $\leq 2n^{1/2-\beta}$ if n is sufficiently large. Lemma 4, applied with $m = 2n^{1/2-\beta}$ and $\varepsilon = \varepsilon'/2$, implies that for all ω from a set of measure $1 - o(1)$ the ε' -neighbourhood of ω meets at least $O(n^{1/4-\beta/2}) = d^{\omega(1)}$ bins. Since accessing each bin requires at least one operation (let even to check that a bin is empty), the theorem is proved. \square

Combining our Theorem 4 with Theorem 2 of Goldberg and Jerrum shows that for all practical purposes the worst-case average performance of metric trees is superpolynomial in dimension of the domain.

Theorem 5 *Let the domain $\Omega = \mathbb{R}^d$ be equipped with a probability measure μ_d in such a way that (\mathbb{R}^d, μ_d) form a normal Lévy family and the μ_d -expected value of the Euclidean distance is $\Theta(1)$. Let \mathcal{F}_d denote a class of functions $f(\theta, x)$ on \mathbb{R}^d parametrized with θ taking values in a space $\mathbb{R}^{\text{poly}(d)}$ and such that computing each value $f(\theta, x)$ takes $d^{O(1)}$ operations of the type described in Thm. 2. Let X be an i.i.d. random sample of \mathbb{R}^d according to μ_d , having n points, where $d = n^{o(1)}$ and $d = \omega(\log n)$. Then, with confidence asymptotically approaching 1, an optimal metric tree indexing scheme for the similarity workload (Ω, ρ, X) whose decision functions belong to the parametrized class \mathcal{F} has expected average performance $d^{\omega(1)}$. \square*

Three remarks are in order to explain the strength of the above result.

(1) Measures μ_d satisfying the above assumption include, for instance, the gaussian distribution, the uniform measure on the unit ball, on the unit sphere, on the unit cube, etc.

(2) A polynomial upper bound on the size of the parameter θ for \mathcal{F} is dictated by the obvious restriction that reading off a parameter of superpolynomial length leads to a superpolynomial lower bound on the length of computation.

(3) In the situations of interest, one can verify that the expected number of datapoints $x \in X$ contained in the smallest query ball meeting X is $O(1)$. For continuous measures on \mathbb{R}^n such as the gaussian measure or the uniform measure on the cube etc., this will be obviously 1. For the Hamming cube, the upper limit of this number as $d \rightarrow \infty$ is bounded by $e \approx 2.7182\dots$. Thus, the lower bound does not come from the fact that there are simply too many valid near neighbours.

Discussion: limitations of the method

The approach to obtaining lower bounds on performance of indexing schemes adopted in this paper consists in combining simple concentration of measure considerations with the basic techniques of statistical learning (VC theory). The argument is applicable to the situation of the following kind. Let $W = (\Omega, \rho, X)$ denote a similarity workload. An indexing scheme for

W consists of a family of real-valued 1-Lipschitz functions f_i , $i \in I$ on Ω , which are in general partially defined: $\text{dom}(f_i) \subseteq \Omega$. Given a query (ω, ε) , where $\omega \in \Omega$ and $\varepsilon > 0$, the algorithm chooses recursively a sequence of indices i_n , based on the previous values $f_{i_k}(\omega)$, $k < n$. At some point, the computation is terminated, and the values $f_{i_k}(\omega)$ point at a collection of bins, whose contents are read off. The role of the functions f_i is to discard those datapoints (or the entire bins) which cannot possibly answer the query. Namely, if $|f_i(\omega) - f_i(x)| \geq \varepsilon$, then, since f_i is a 1-Lipschitz function, one has $d(\omega, x) \geq \varepsilon$, and so the point x is irrelevant. All the points (or entire bins) which cannot be discarded are returned and their contents checked against the condition $d(x, \omega) < \varepsilon$.

On the spaces of high dimension, every 1-Lipschitz function concentrates sharply near its mean (or median) value. If in addition we assume that the class \mathcal{F} of all functions used for a particular indexing scheme has a low complexity in the sense of VC dimension, we can conclude that the number of points discarded by every function f_i drops off fast as dimension d of the domain grows, resulting in degrading performance.

So far, we are aware of essentially two different types of such indexing schemes: metric trees (treated in the present paper) and pivot tables [6]. For pivots, the methods of the present paper have been subsequently used to derive an expected average performance lower bound $O(n/d)$ [40]. It is not clear to the author how to state a more general result from which both estimates would follow, nor whether such a result would be useful in view of lack of other examples.

Even if the cell-probe model has some formal similarities with the metric tree scheme (a hierarchical tree structure, a collection of cells as an indexing scheme, computations performed at each node with a limited number of cells accessed, etc.), it is not clear whether the partially defined functions determined by the algorithm at each node will be 1-Lipschitz (they are taking values in the Hamming cube). The examples of implemented indexing schemes for *exact* nearest neighbour search known to this author seem to be using 1-Lipschitz functions, but of course this does not preclude the existence of schemes based on other ideas.

Furthermore, assuming that an indexing scheme consists of a family of 1-Lipschitz functions whose values are recursively computed by the algorithm does not necessarily imply that the role of the functions is reduced to certifying that a certain point is not in the ε -ball around the query point. As an example, consider the indexing scheme based on a walk on the Delaunay graph of X in Ω and called *spatial approximation* by its author [23]. For every datapoint $x \in X$, the scheme stores a list of datapoints whose Voronoi cells are adjacent to the cell containing x . At the search phase, a sequence of datapoints x_1, x_2, \dots, x_n is chosen, where each x_{i+1} is the closest point to ω on the list of points Delaunay-adjacent to x_i . If choosing x_{i+1} so as to get closer to ω is impossible, one backtracks. In practice, the scheme performs on par with the state of the art pivot or metric tree based schemes [25]. We

do not know whether our methods can be employed to prove the curse of dimensionality for this particular scheme in the same general setting.

It appears that attempting to extend the method to randomized, approximated NN search stands no chance either. Firstly, the dimensionality reduction-type methods often present in randomized algorithms for approximate search [18,16,1] mean that instead of 1-Lipschitz functions, one is using what may be called “probably approximately 1-Lipschitz” ones. For instance, a random projection from a high-dimensional Euclidean space to a subspace of smaller dimension, appropriately rescaled, will have the property that for most pairs of points x, y the distance between them is approximately preserved, to within a factor of $1 \pm \epsilon$. This property in itself is a consequence of concentration of measure, but such maps do not exhibit a strong concentration property, rendering our methods inapplicable.

Chapter 4 in [43] discusses algorithms for approximate similarity search based on a traditional metric tree, equipped with 1-Lipschitz decision functions, but employing aggressive pruning, either randomized or deterministic. Even here, our proof does not seem to be readily transferable. Indeed, it is based on the basic premise that *every bin meeting the ϵ -neighbourhood of the query point needs to be examined in a deterministic fashion*. A randomized algorithm, on the contrary, avoids opening bins which are deemed unlikely to contain relevant datapoints. Experiments confirm that some of the algorithms in question perform up to 300 times faster than the corresponding algorithms for exact search using the same indexing structure (*loc.cit.*), and provide a circumstantial evidence that the situation here is indeed fundamentally different and possibly not amenable to the same methods of analysis.

Where does this leave us? In the opinion of the referee,

“...the present lower bound proof might be too specific to the model (as opposed to the actual problem).”

The first part of the statement, at least for the time being, appears to be well-founded. It is however less clear whether the curse of dimensionality conjecture for the Hamming cube is the only theoretical lens possible through which to regard the “the actual problem” of the curse of dimensionality. The model studied in the present article covers a rather wide class of popular indexing schemes and obtains superpolynomial lower bounds on their performance within mathematically exacting standards of statistical learning. While the setting of artificially high-dimensional synthetic i.i.d. data fed to the scheme is not realistic either, our results still provide a better insight in how these particular schemes do function.

Indeed, most data practitioners seem to believe that the intrinsic dimension of real-life datasets does not exceed as few as perhaps seven or ten dimensions. A deeper understanding of underlying geometry of workloads and its interplay with complexity is called for in order to learn to detect and use this low dimensionality efficiently, and asymptotic analysis of algorithm

performance in an artificial setting of very high dimensions is contributing towards this goal.

References

1. A. Andoni, P. Indyk, M. Pătrascu, *On the optimality of the dimensionality reduction method*, in: Proc. 47th IEEE Symp. on Foundations of Computer Science, pp. 449–458, 2006.
2. M. Anthony and P. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999.
3. O. Barkol and Y. Rabani. Tighter lower bounds for nearest neighbor search and related problems in the cell probe model. In: *Proc. 32nd ACM Symp. on the Theory of Computing*, 2000, pp. 388–396.
4. K. Beyer, J. Goldstein, R. Ramakrishnan, and U. Shaft. *When is “nearest neighbor” meaningful?*, in: *Proc. 7-th Intern. Conf. on Database Theory (ICDT-99)*, Jerusalem, pp. 217–235, 1999.
5. A. Borodin, R. Ostrovsky, and Y. Rabani. Lower bounds for high-dimensional nearest neighbor search and related problems, in: *Proc. 31st Annual ACS Sympos. Theory Comput.*, 312–321, 1999.
6. Bustos, B., Navarro, G., Chávez, E. (2003) Pivot selection techniques for proximity searching in metric spaces. *Pattern Recognition Lett.*, vol. 24, pp. 2357–2366.
7. E. Chávez, G. Navarro. *A compact space decomposition for effective metric indexing*. *Pattern Recognition Letters* 26:1363–1376, 2005.
8. E. Chávez, G. Navarro, R. Baeza-Yates and J. L. Marroquín. Searching in metric spaces. *ACM Computing Surveys* 33:273–321, 2001.
9. P. Ciaccia, M. Patella and P. Zezula. M-tree: An efficient access method for similarity search in metric spaces. In *Proceedings of 23rd International Conference on Very Large Data Bases (VLDB’97)*, (Athens, Greece), 426–435, 1997.
10. P. Ciaccia, M. Patella and P. Zezula. A cost model for similarity queries in metric spaces, in: *Proc. 17-th ACM Symposium on Principles of Database Systems (PODS’98)*, Seattle, WA, 59–68, 1998.
11. A. Faragó, T. Linder, and G. Lugosi, Fast nearest neighbor search in dissimilarity spaces, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 18, pp. 957–962, 1993.
12. P.W. Goldberg and M.R. Jerrum, *Bounding the Vapnik–Chervonenkis dimension of concept classes parametrized by real numbers*, Machine Learning 18:131–148, 1995.
13. Gromov, M., Milman, V.D. (1983). A topological application of the isoperimetric inequality. *Amer. J. Math.* 105, 843–854.
14. J. M. Hellerstein, E. Koutsoupias, D. P. Miranker, C. Papadimitriou, and V. Samoladas. On a model of indexability and its bounds for range queries. *Journal of the ACM (JACM)*, 49(1):35–55, 2002.
15. P. Indyk. Nearest neighbours in high-dimensional spaces. In: J.E. Goodman, J. O’Rourke, Eds., *Handbook of Discrete and Computational Geometry*, Chapman and Hall/CRC, Boca Raton–London–New York–Washington, D.C. 877–892, 2004.

16. Piotr Indyk, Rajeev Motwani, *Approximate nearest neighbors: towards removing the curse of dimensionality*, Proceedings of the thirtieth annual ACM symposium on Theory of computing, pp. 604–613, 1998, Dallas, Texas.
17. R. Krauthgamer and J.R. Lee, *The black-box complexity of nearest-neighbor search*, Theoretical Computer Science 348(2):262 - 276, December 2005.
18. E. Kushilevitz, R. Ostrovsky, Y. Rabani, Efficient Search for Approximate Nearest Neighbor in High Dimensional Spaces. SIAM Journal on Computing 30:457–474, 2000.
19. M. Ledoux. *The Concentration of Measure Phenomenon*, volume 89 of *Mathematical Surveys and Monographs*. American Mathematical Society, Providence, RI, 2001.
20. S. Mendelson, A few notes on statistical learning theory. In: S. Mendelson, A.J. Smola, Eds., *Advanced Lectures in Machine Learning*, LNCS 2600, pp. 1–40, Springer, 2003.
21. V.D. Milman and G. Schechtman, *Asymptotic Theory of Finite Dimensional Normed Spaces*, volume 1200 of *Lecture Notes in Mathematics*. Springer, 1986.
22. P.B. Miltersen, *Cell probe complexity - a survey*. In: 19th Conference on the Foundations of Software Technology and Theoretical Computer Science (FSTTCS), 1999. Advances in Data Structures Workshop.
23. Gonzalo Navarro, *Searching in metric spaces by spatial approximation*, The VLDB Journal 11:28–46, August 2002.
24. Gonzalo Navarro, *Analysing metric space indexes: what for?* Invited paper, in: Proc. 2nd Int. Workshop on Similarity Search and Applications (SISAP 2009), Prague, Czech Republic, 2009, 3–10.
25. Gonzalo Navarro, Nora Reyes, *Dynamic spatial approximation trees for massive data*, in: Proc. 2nd Int. Workshop on Similarity Search and Applications (SISAP 2009), Prague, Czech Republic, 2009, pp. 81–88.
26. R. Panigrahy, K. Talwar, U. Wieder, *A geometric approach to lower bounds for approximate near-neighbor search and partial match*, in: Proc. 49th IEEE Symp. on Foundations of Computer Science, pp. 414–423, 2008.
27. M. Patrascu, M. Thorup, *Higher lower bounds for near-neighbor and further rich problems*, in Proc. 47th IEEE Symp. on Foundations of Computer Science, pp. 646–654, 2006.
28. V. Pestov. On the geometry of similarity search: dimensionality curse and concentration of measure. *Inform. Process. Lett.*, 73:47–51, 2000.
29. V. Pestov, Intrinsic dimension of a dataset: what properties does one expect? in: *Proc. of the 22-nd Int. Joint Conf. on Neural Networks (IJCNN'07)*, Orlando, FL, pp. 1775–1780, 2007.
30. V. Pestov. An axiomatic approach to intrinsic dimension of a dataset. *Neural Networks*, 21:204–213, 2008.
31. V. Pestov and A. Stojmirović. Indexing schemes for similarity search: an illustrated paradigm. *Fund. Inform.*, 70:367–385, 2006.
32. H. Samet. *Foundations of Multidimensional and Metric Data Structures*. Morgan Kaufmann Publishers Inc., San Francisco, CA, 2005.
33. S. Santini, *Exploratory Image Databases: Content-Based Retrieval*, Academic Press, Inc. Duluth, MN, USA, 2001.
34. U. Shaft and R. Ramakrishnan. Theory of nearest neighbors indexability. *ACM Transactions on Database Systems (TODS)*, 31:814–838, 2006.

35. A. Stojmirović and V. Pestov. Indexing schemes for similarity search in datasets of short protein fragments. *Information Systems*, 32:1145–1165, 2007.
36. J.K. Uhlmann. Satisfying general proximity/similarity queries with metric trees, *Information Processing Letters* 40:175–179, 1991.
37. V.N. Vapnik. *Statistical Learning Theory*. John Wiley & Sons, Inc., New York, 1998.
38. S.S. Vempala. *The Random Projection Method*. DIMACS Series in Discrete Mathematics and Theoretical Computer Science, **65**, Amer. Math. Soc., Providence, R.I., 2004.
39. M. Vidyasagar. *Learning and Generalization, With Applications to Neural Networks*. Second Ed. Springer-Verlag, London, 2003.
40. I. Volnyansky and V. Pestov, *Curse of dimensionality in pivot-based indexes*. - Proc. 2nd Int. Workshop on Similarity Search and Applications (SISAP 2009), Prague, Czech Republic, 2009, pp. 39-46.
41. R. Weber, H.-J. Schek, and S. Blott, A quantitative analysis and performance study for similarity-search methods in high-dimensional spaces. in: *Proceedings of the 24-th VLDB Conference*, New York, pp. 194–205, 1998.
42. P. Yianilos. Data structures and algorithms for nearest neighbor search in general metric spaces, in: *Proc. 3rd Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 311–321, 1993.
43. P. Zezula, G. Amato, Y. Dohnal, and M. Batko. *Similarity Search. The Metric Space Approach*. Springer Science + Business Media, New York, 2006.